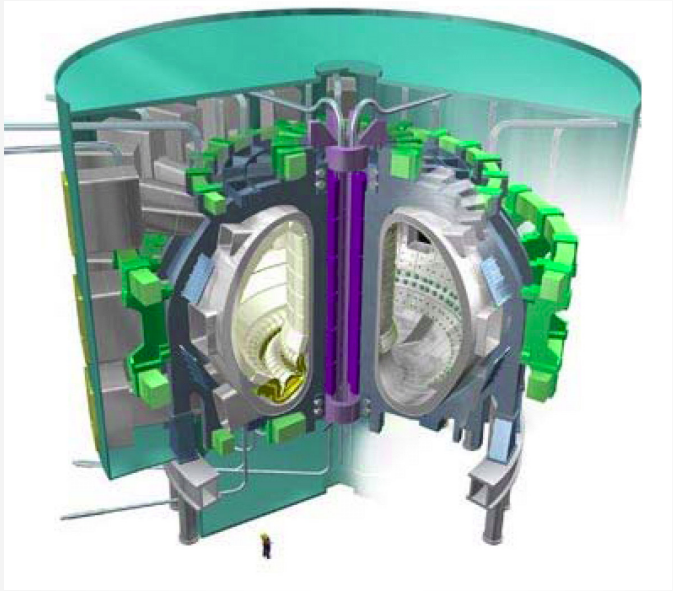
Towards Petascale Computing in Support of ITER

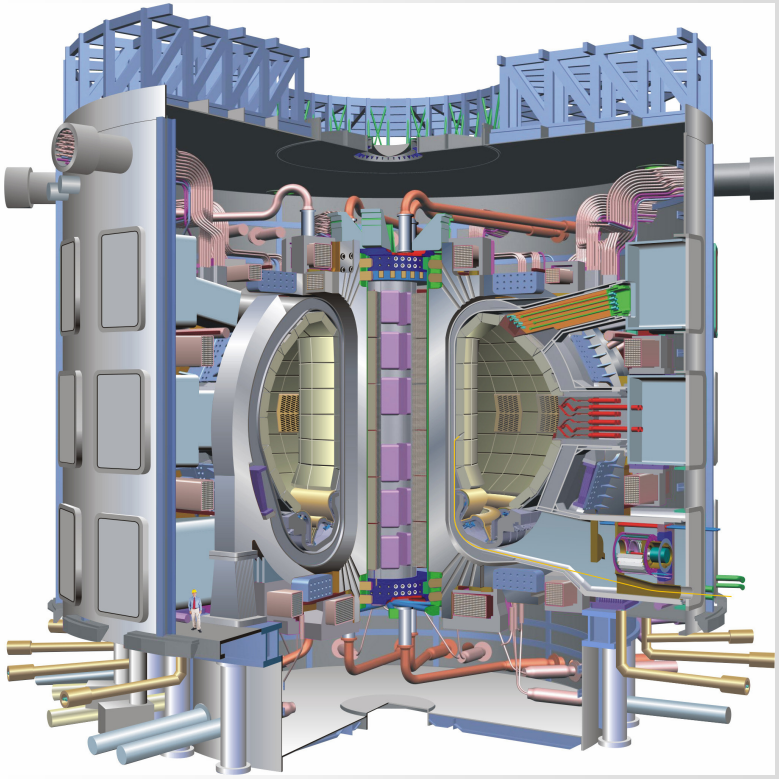


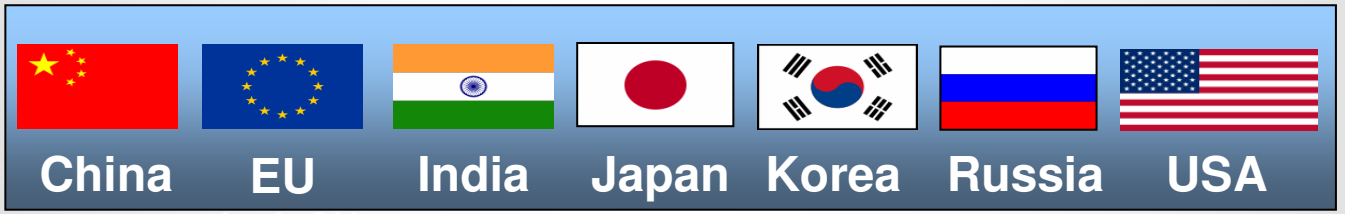
H. Lederer, R. Tisma, R. Hatzky, A. Bottino*, F. Jenko*

Garching Computing Center of the Max Planck Society
*Max Planck Institute for Plasma Physics
D-85748 Garching, Germany

ITER Experiment

www.iter.org





incl. CH

RZG IPP

Theory Support for ITER

- Large scale numerical simulations will be a necessity.
- Plasma turbulence simulations play a key role for the design, construction and optimization of the necessary fusion devices.
- The simulations will be compute and memory intensive
- Applications must be able to efficiently use tens of thousands of processors.
- Highly scalable applications are mandatory.

Important European Simulation Codes for Plasma Core Turbulence with ITER relevance

- **ORB5 (CRPP, Lausanne & IPP, Greifswald & Garching)**
- **GENE (IPP, Garching)**
- **GYSELA (CEA, Cadarache)**

Simulation Code ORB5

The ORB code family uses a particle-in-cell (PIC), time evolution approach, and takes advantage of all the recent techniques of noise reduction and control in PIC simulations.

ORB uses a statistical optimisation technique that increases the accuracy by orders of magnitude.

Initiated at CRPP (Centre de Recherches en Physique des Plasmas), Lausanne, ORB has been substantially upgraded at MPI for Plasma Physics (IPP), Garching.

Ongoing code development is made under a close collaborative effort between IPP and CRPP

Simulation Code GENE

GENE: so-called continuum (or Vlasov) code.

All differential operators in phase space are discretized via a combination of spectral and higher-order finite difference methods.

For maximum efficiency, GENE uses a coordinate system which is aligned to the equilibrium magnetic field and a reduced (flux-tube) simulation domain.
This reduces the computational effort by 2-3 orders of magnitude.

GENE can deal with arbitrary toroidal geometry (tokamaks or stellarators) and retains full ion/electron dynamics as well as magnetic field fluctuations.

At present, GENE is the only plasma turbulence code in Europe with such capabilities.

ORB5 Code: Single processor optimization

Two bottlenecks identified and improved.

- **implementation of the FFT** - **cache sort of the Monte Carlo particles**

FFT: module written with different interfaces to specialized FFT libraries.
(original code included FFT source code with poor performance)

Interfaces to FFTs from IBM ESSL, Intel MKL, FFTW
-> no restrictions to vector lengths of powers of two
-> more flexibility to choose the grid resolution of the electrostatic potential

Cache sort:

- sorting the Monte Carlo particles relative to their position in the grid cells of the electrostatic potential, results in high cache reuse of the electrostatic field sampling
- overhead caused by introduction of the sort routine was minimized
- option to enlarge a work array for the sorting process (can speed up the isort routine by a factor of three)

ORB5 Code: Improving scalability

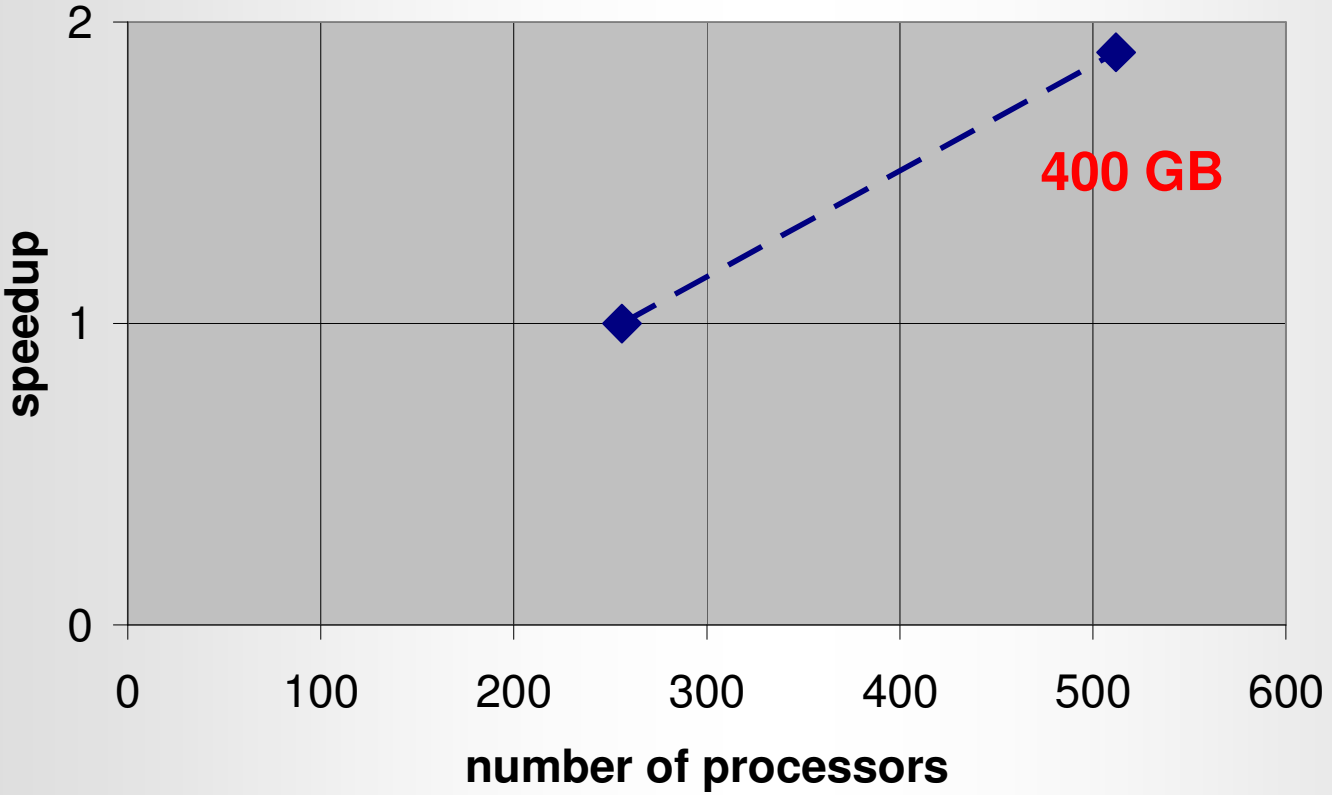
Implementation of the domain cloning concept:
- optimizes scaling
- decouple the selectable grid resolution from the number of processors used for the simulation.

Domain cloning builds on two decomposition techniques:

1. For the domain decomposition, different portions of the physical domain and of the corresponding electrostatic field grid are assigned to different processors together with the Monte Carlo particles that reside on them. As particles move from one region to another, they are communicated to the processor which is associated with the new region.

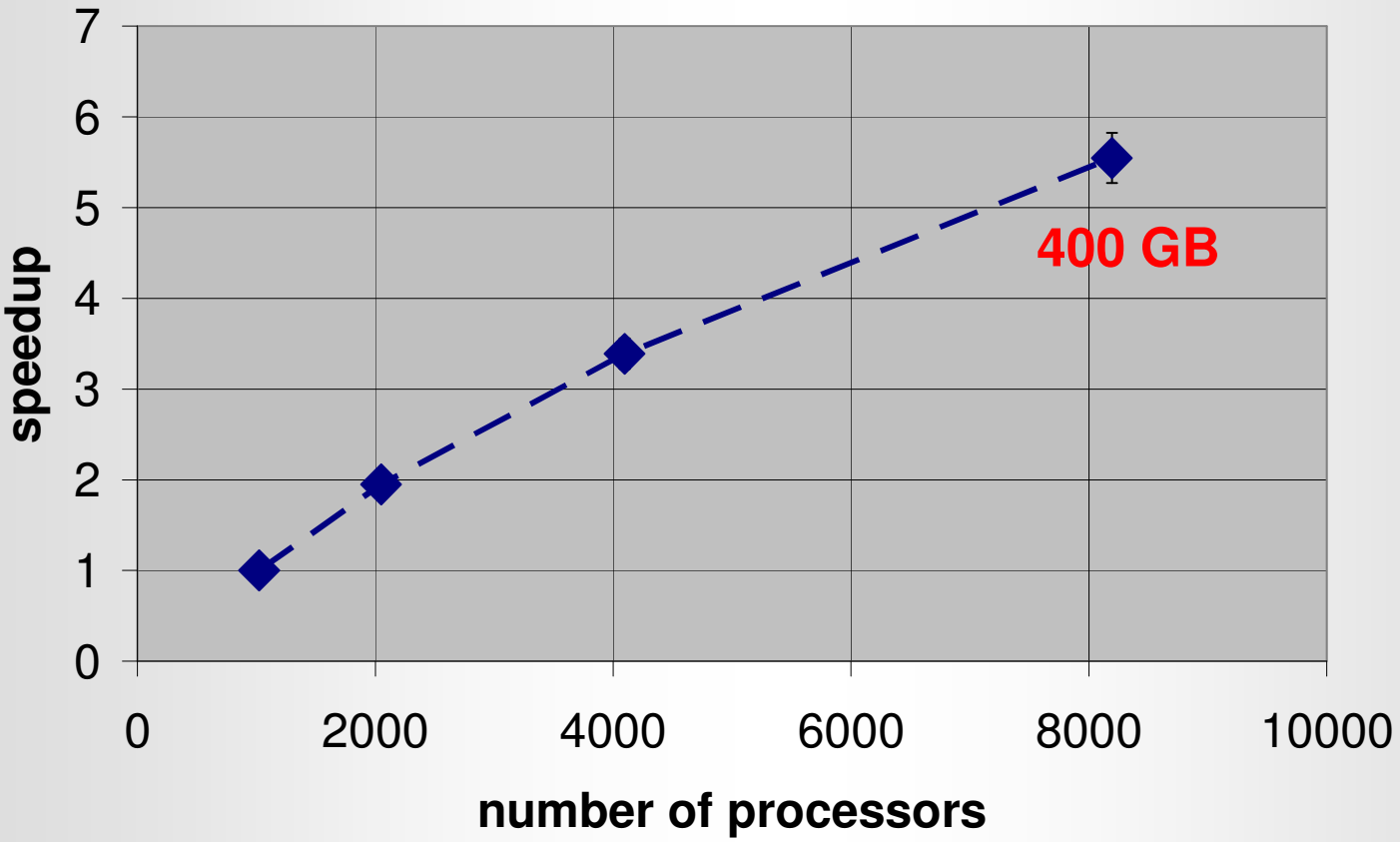
2. Particle decomposition: the whole spatial grid is assigned to every processor, but each processor takes care of only a subset of the particle population. Partial contributions to the ion density, which is required to update the electrostatic field, are communicated among processors and summed via global sum operations.

ORB5 Code: Test of scalability on IBM Power4



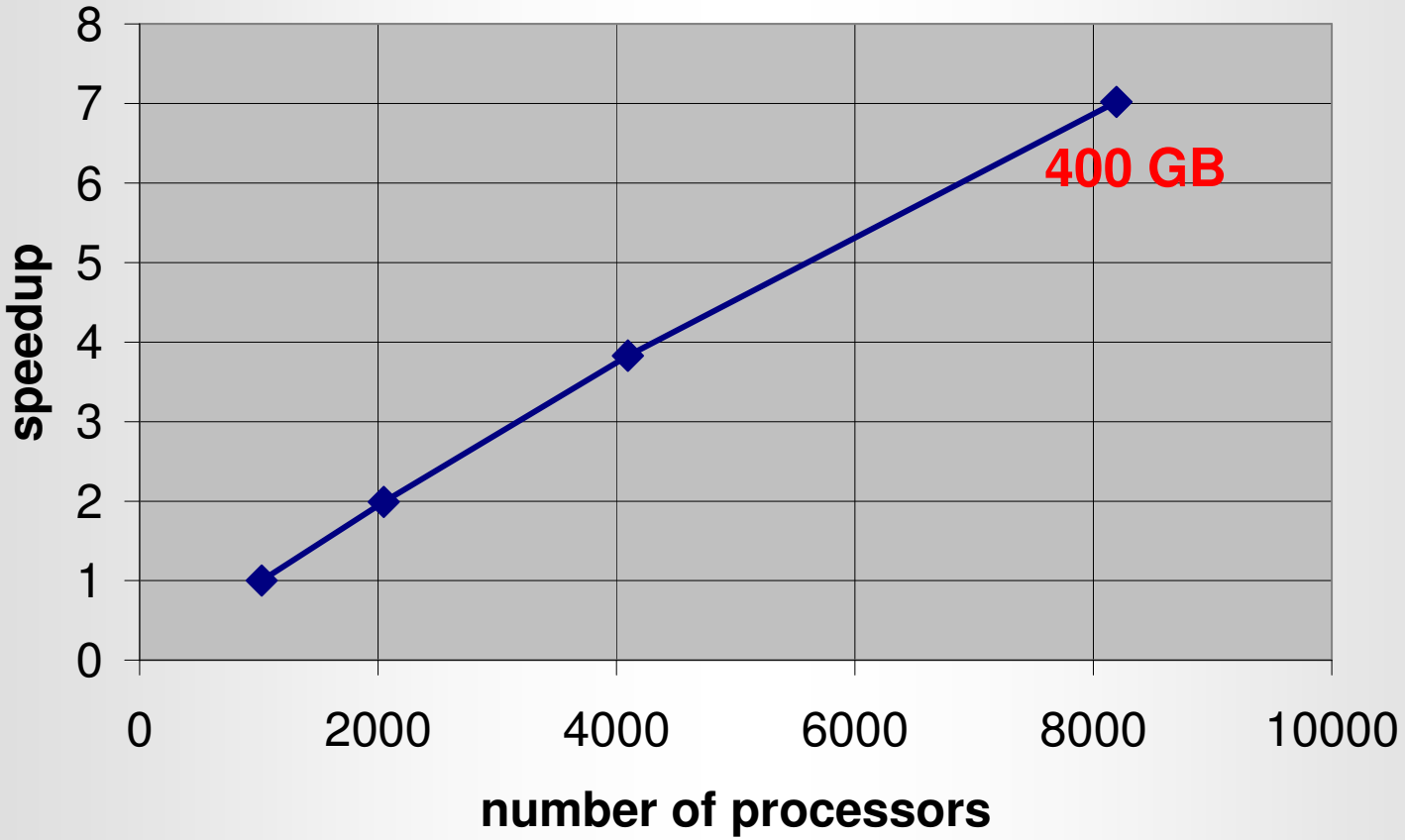
Strong scaling of ORB5 for an ITG simulation (~400 GB)
512 proc. result normalized on the 256 proc. result.
(Measurements on IBM p690@1.3 GHz + HPS at RZG)

ORB5 Code: Scalability test on Cray XT3



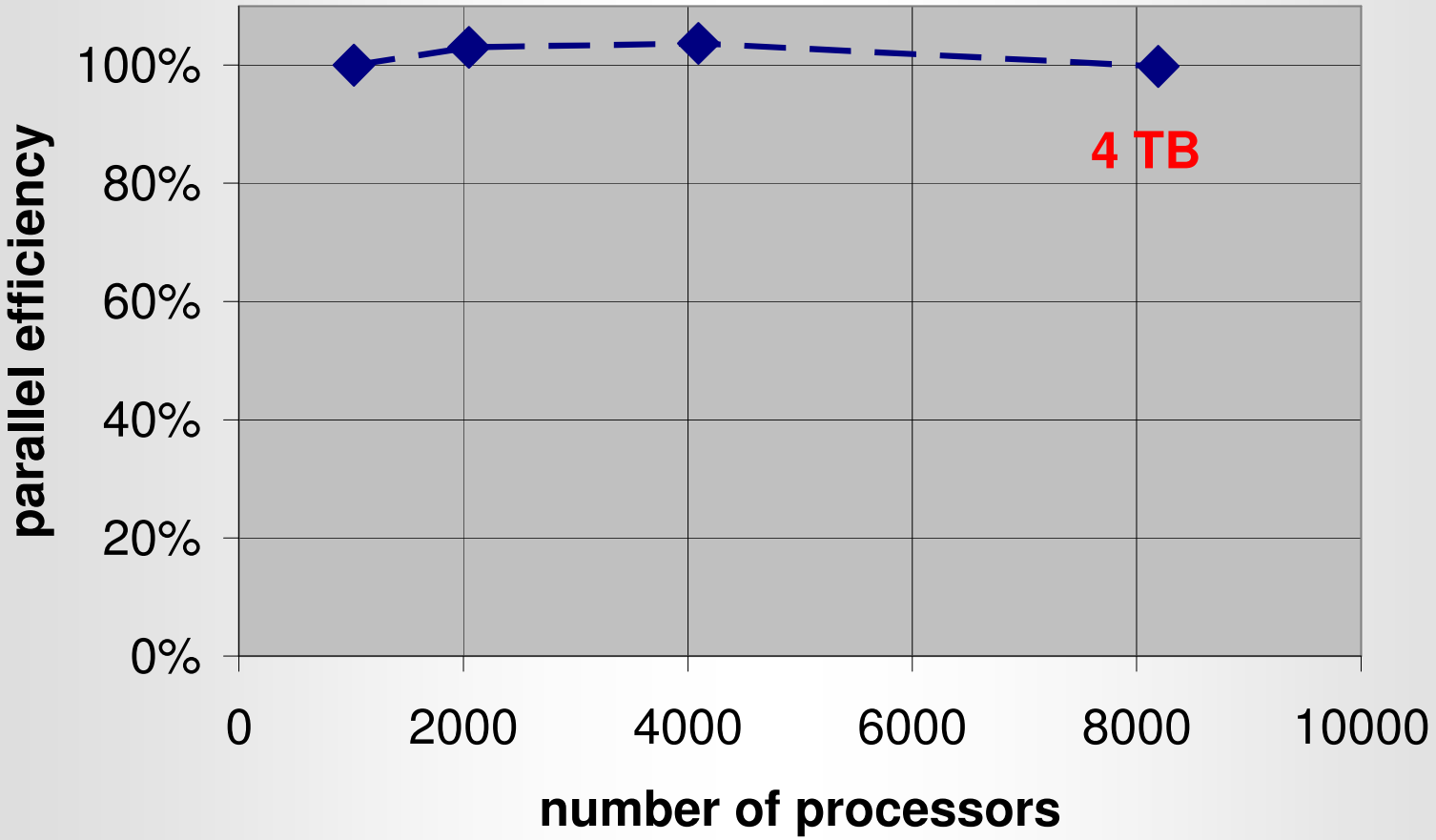
Strong scaling of ORB5 for an **ITG** simulation (0.5 billion particles)
(results normalized on the result for 1024 processor-cores)
(Measurements on Cray XT3 at ORNL; courtesy of Cray Inc.)

ORB5 Code: Scalability test on BlueGene/L



Strong scaling of ORB5 for an **ETG** simulation (0.8 billion particles)
Results normalized on the 1024 processor result
(Measurements on BlueGene/L at IBM Watson Research Center in co-processor mode)

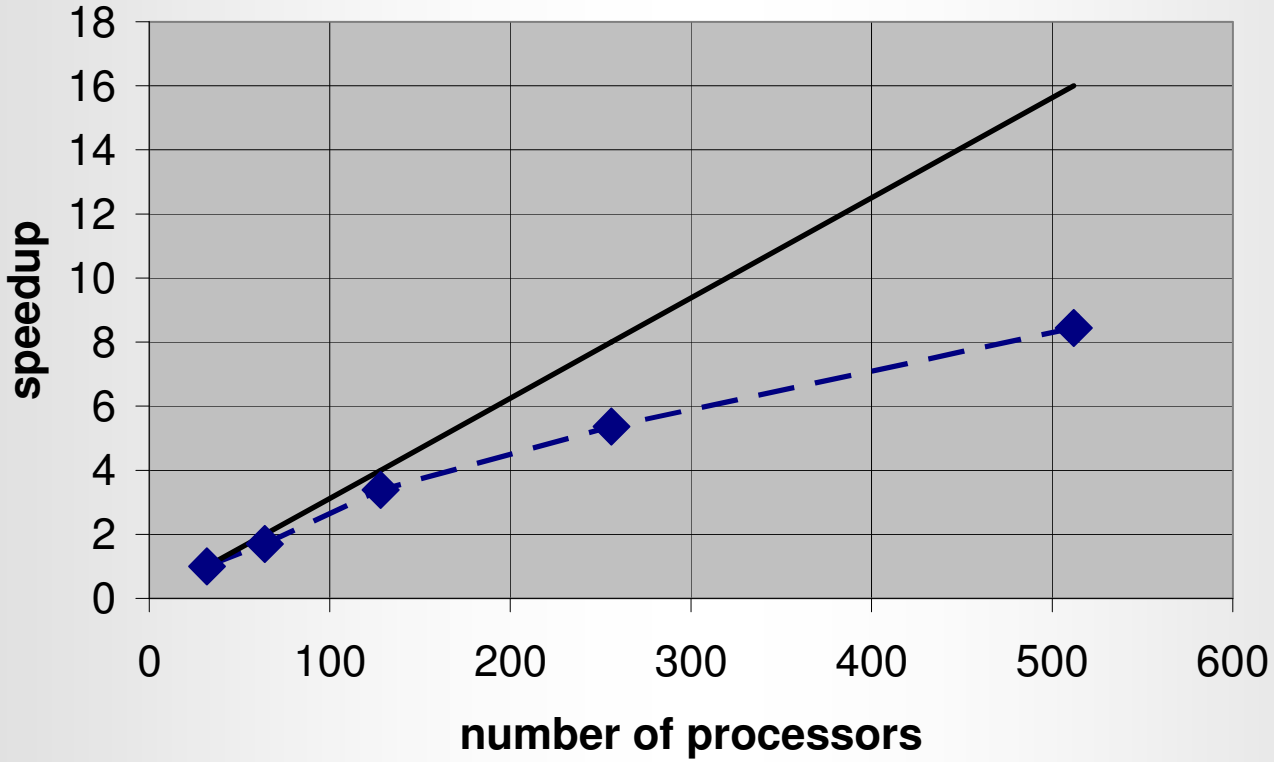
ORB5 Code: Scalability test on BlueGene/L



Weak scaling of ORB5 for **ETG** simulation (~ 0.8 M particles per processor)
Results normalized on the 1024 processor result
(Measurements on BlueGene/L at IBM Watson Research Center in co-processor mode)

Analysis of GENE v 9

- **Mixed parallelization model MPI+OpenMP**
- **Upper limit of 64 MPI tasks**



-> Limited scalability

Analysis of GENE v 9

6 dimensions available for parallelisation:
species + 3 space coordinates. + 2 velocity coordinates

The spatial coordinates *x* and *y*, in GENE v 9 only treated serially, contain significant potential for domain decomposition.

A large number of 2-dimensional FFTs done on the *xy* planes. If the *xy* plane is distributed, it must be transposed in order to perform the FFT in the *x* and *y* directions.

However, the transposition requires an all-to-all communication
⇨ high communication overhead.

If the y-coordinate could be left in k-space, the FFTs would have to be performed on the *x*-coordinate only, which is contiguous in memory. A transpose of the *xy* plane would **not** be necessary.

Improving the scalability of GENE

- Sections in the code where transformations to the configuration space are necessary.

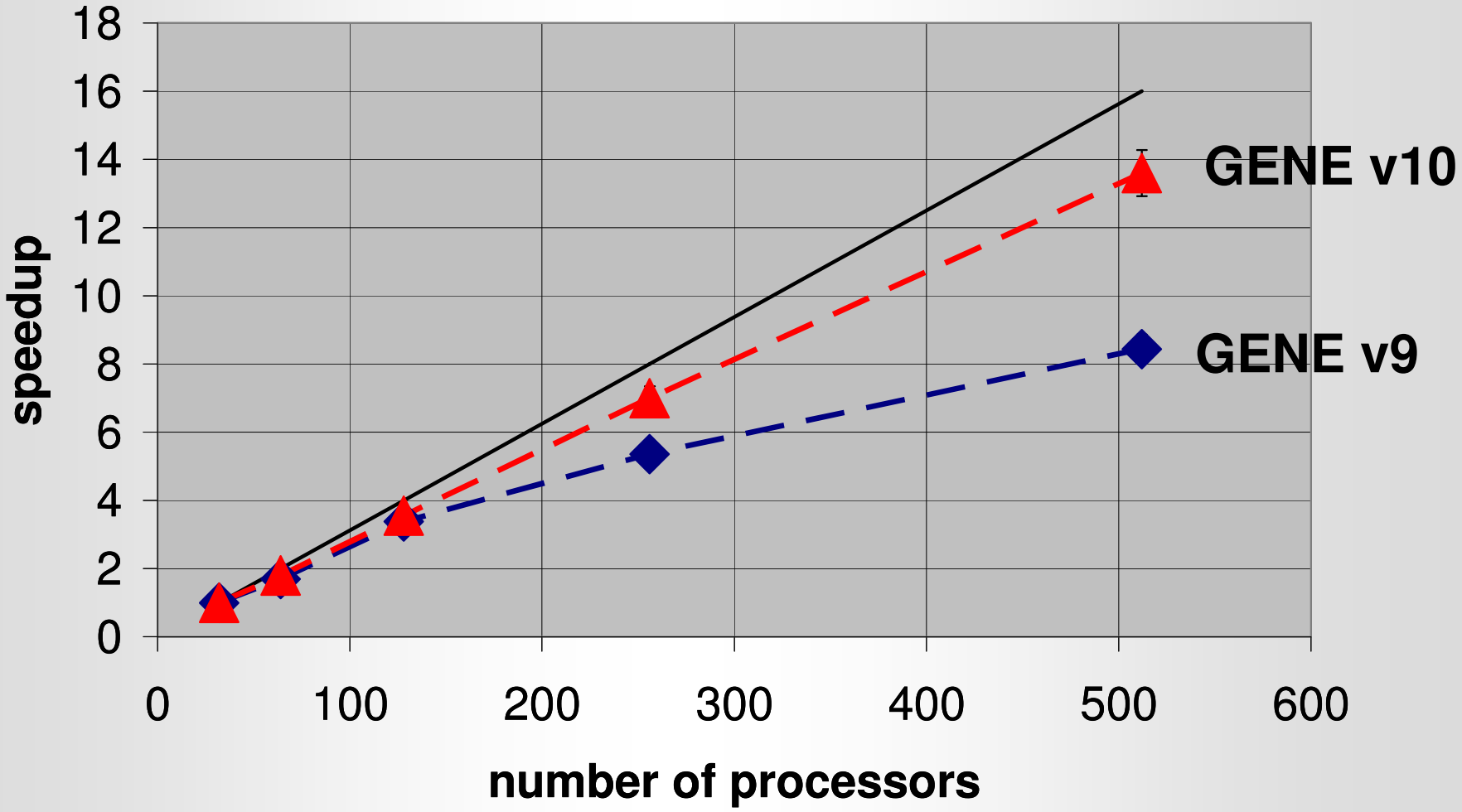
- For these transformations, transpositions of the *xy* plane have to be performed. However, the number of these transformations is much smaller than the number of transformations to k-space.

- Implication of a major change to the overall data structure
  -> consequences for many parts of the code.

- Prerequisite: code adaptation by the authors according to the new role of the *y*-coordinate.

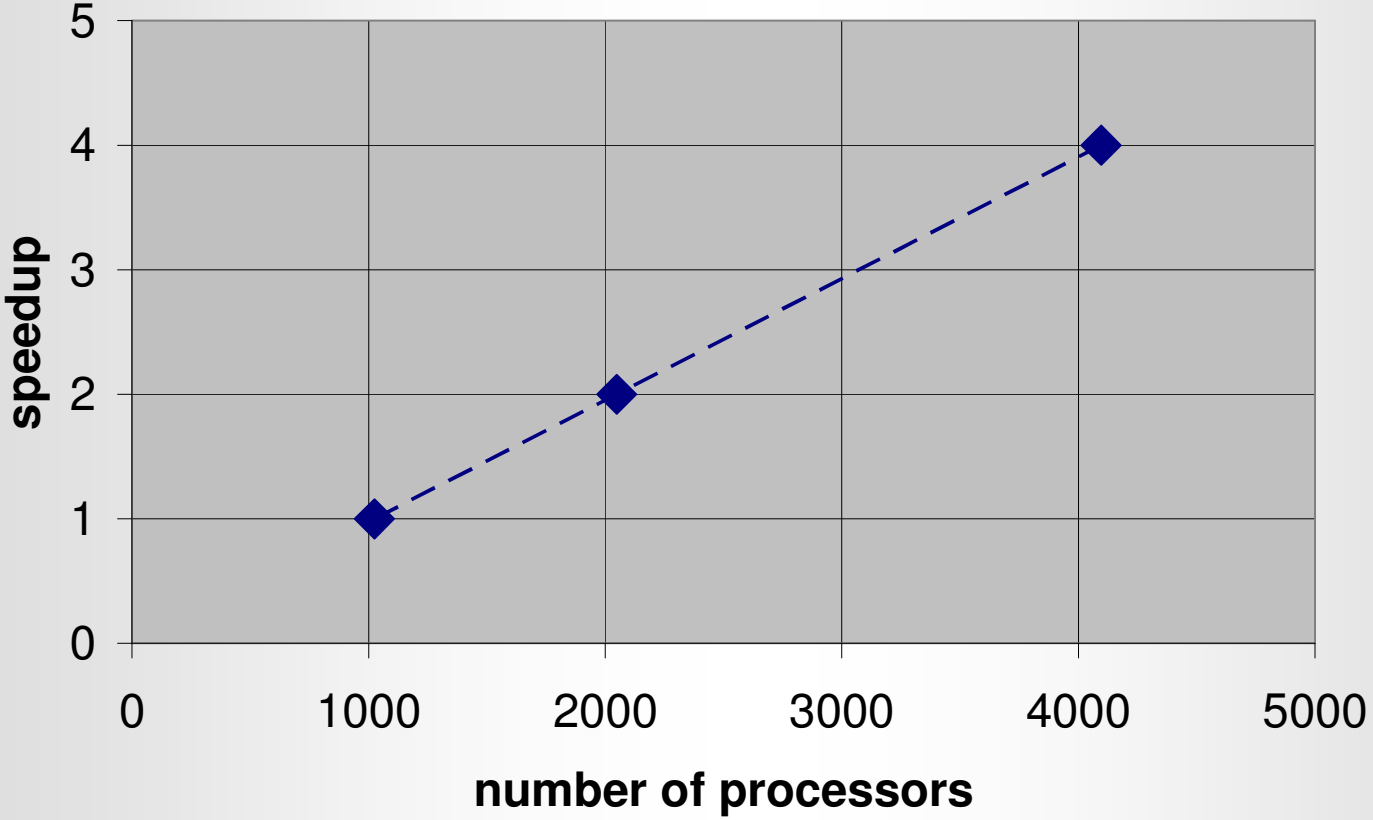
- Main task: design and realization of the domain decomposition of the *y*-coordinate. Dynamical mapping of the number of points available in the *y* direction on the number of processors selected for treating the *y* direction (consequently applied to all loops in the *y* direction (~ 20 to 40).

GENE scalability on IBM p690 at RZG



RZG IPP

Verification of GENE v10 scalability on a larger number of processors on Cray XT3



Strong scaling of GENEv10 (problem size of ~300-500 GB)
normalized to 1024 processors (Measurements courtesy of Cray Inc.)

Enhancing portability of GENE

FFT-routines: Interfaces to

- IBM proprietary ESSL-library (on IBM systems)
- Math Kernel Library (MKL) from INTEL (on SGI Altix)
- FFTW package (on Cray XT3/4)

Bessel functions:

Routines originally used from the NAG library were replaced: three new subroutines were written implementing these Bessel functions

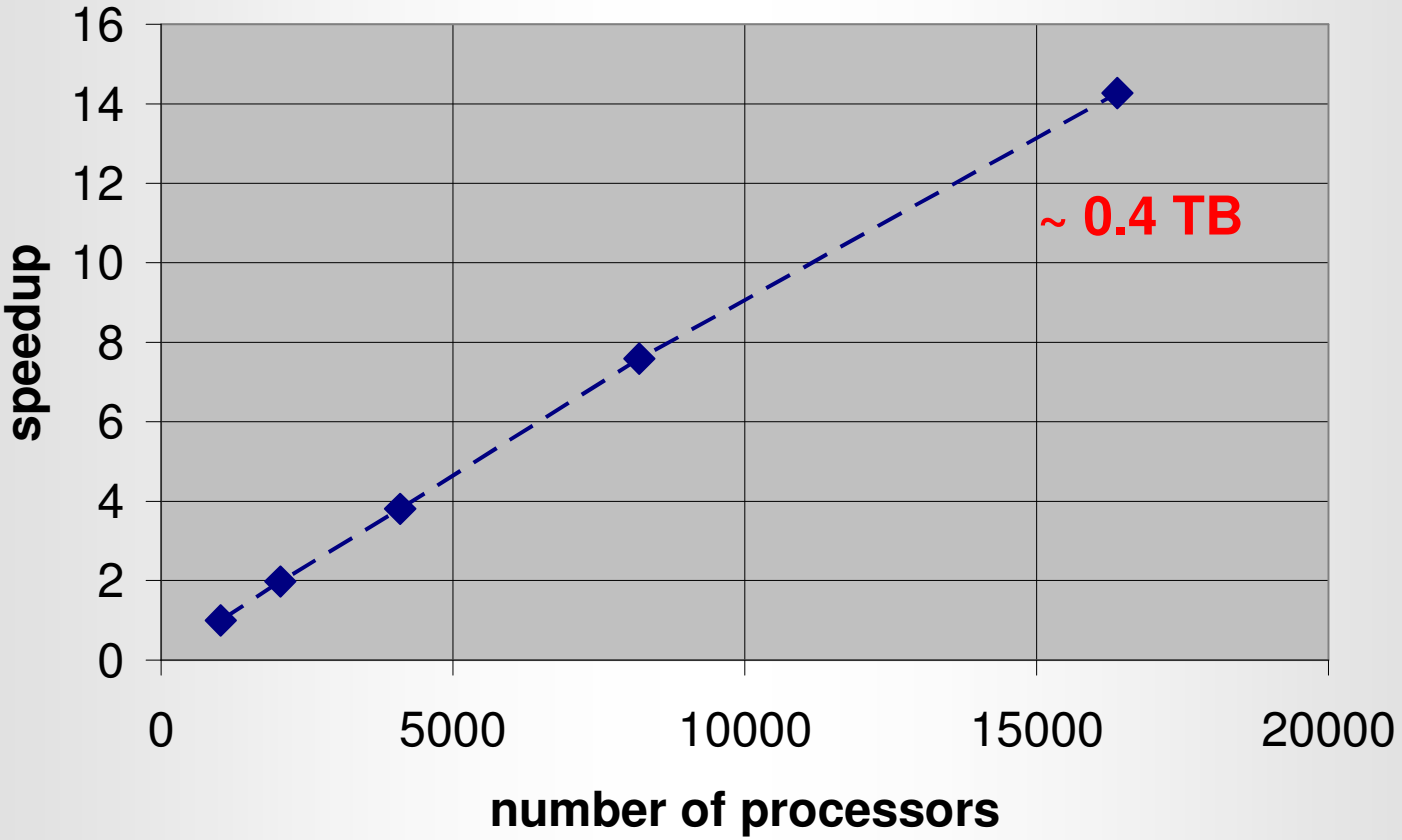
GENE v11
and tests on higher processor numbers

- GENE v11: Enhancements of algorithms and of functionality by authors
- Parallelization scheme of GENEv11 same as that of GENEv10
- Tests of GENE v11 scalability on higher processor numbers
- Porting of GENE v11 to IBM BlueGene/L
- Test of GENE v11 up to 8k processors on BG/L
- Results on BG/L for strong scaling measurements
  - on 4k processors: quasi-linear speedup
  - on 8k processors: degradation (parallel efficiency: 73%)

Further improvements of scalability?

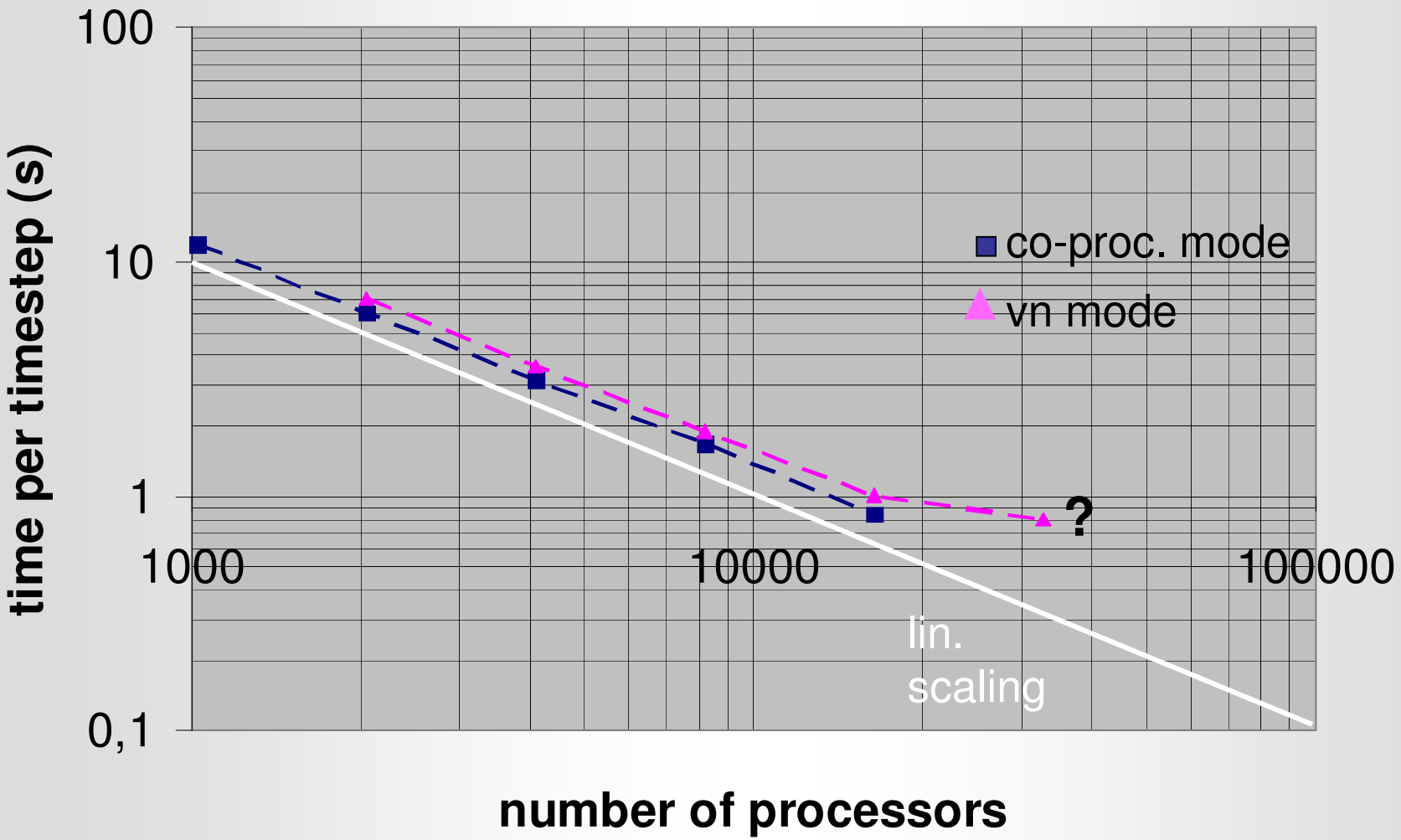
- Still potential for parallelization: *z* velocity dimension
- Parallelization of the *z* dimension by domain decomposition
- New GENEv11+
- Test of GENEv11+ on BG/L
- Results on BG/L for strong scaling measurements:
  - on 4k processors: quasi-linear speedup
  - on 8k processors: quasi-linear speedup (parallel efficiency: 95%)

GENE v11+ on IBM BlueGene/L



Strong scaling of GENEv11+ normalized to 1k processors
(problem ~300-500 GB; measurements in co-processor mode at IBM Watson Research Center)

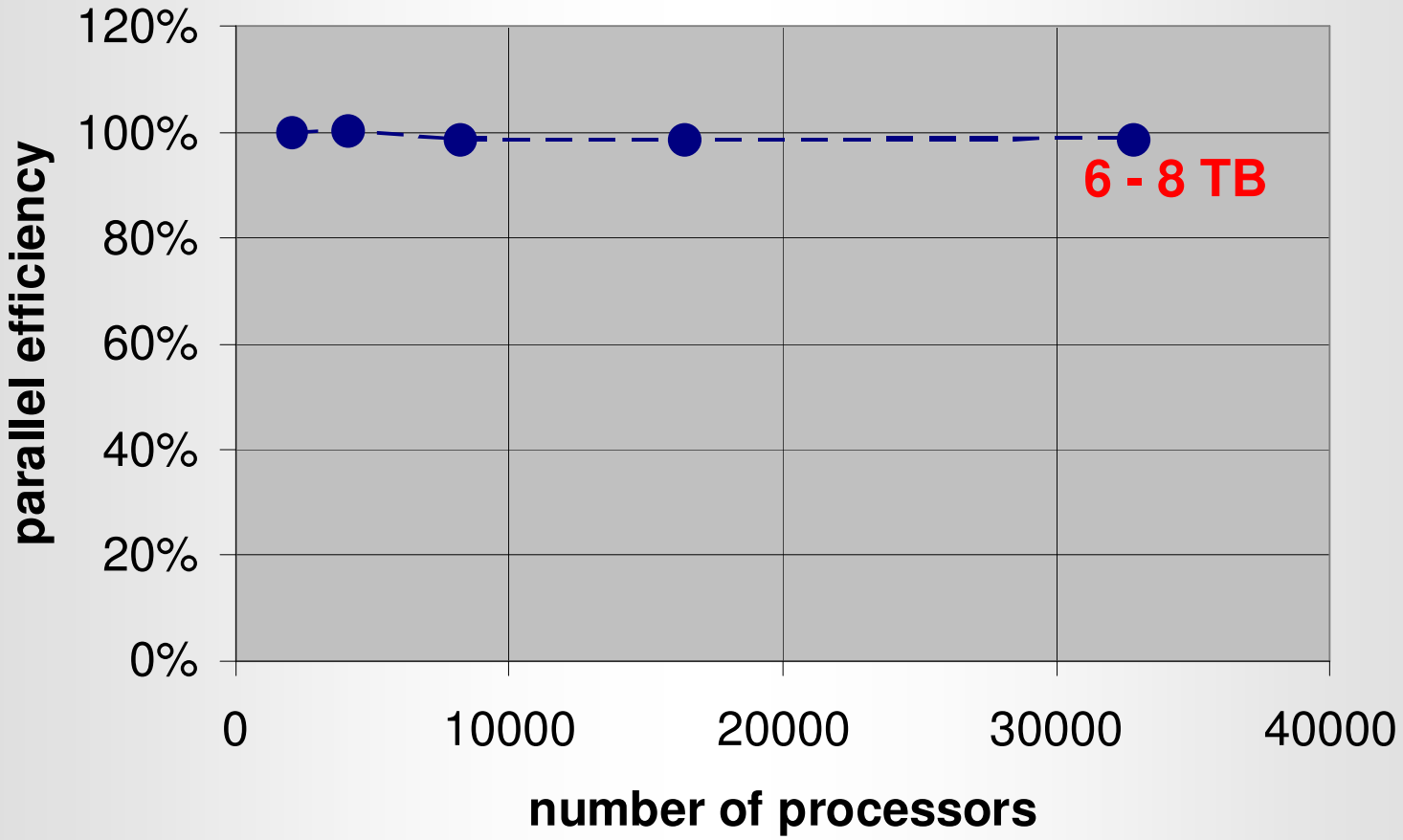
GENE v11+ on IBM BlueGene/L



Strong scaling of GENEv11+ (problem ~400 GB)

RZG IPP

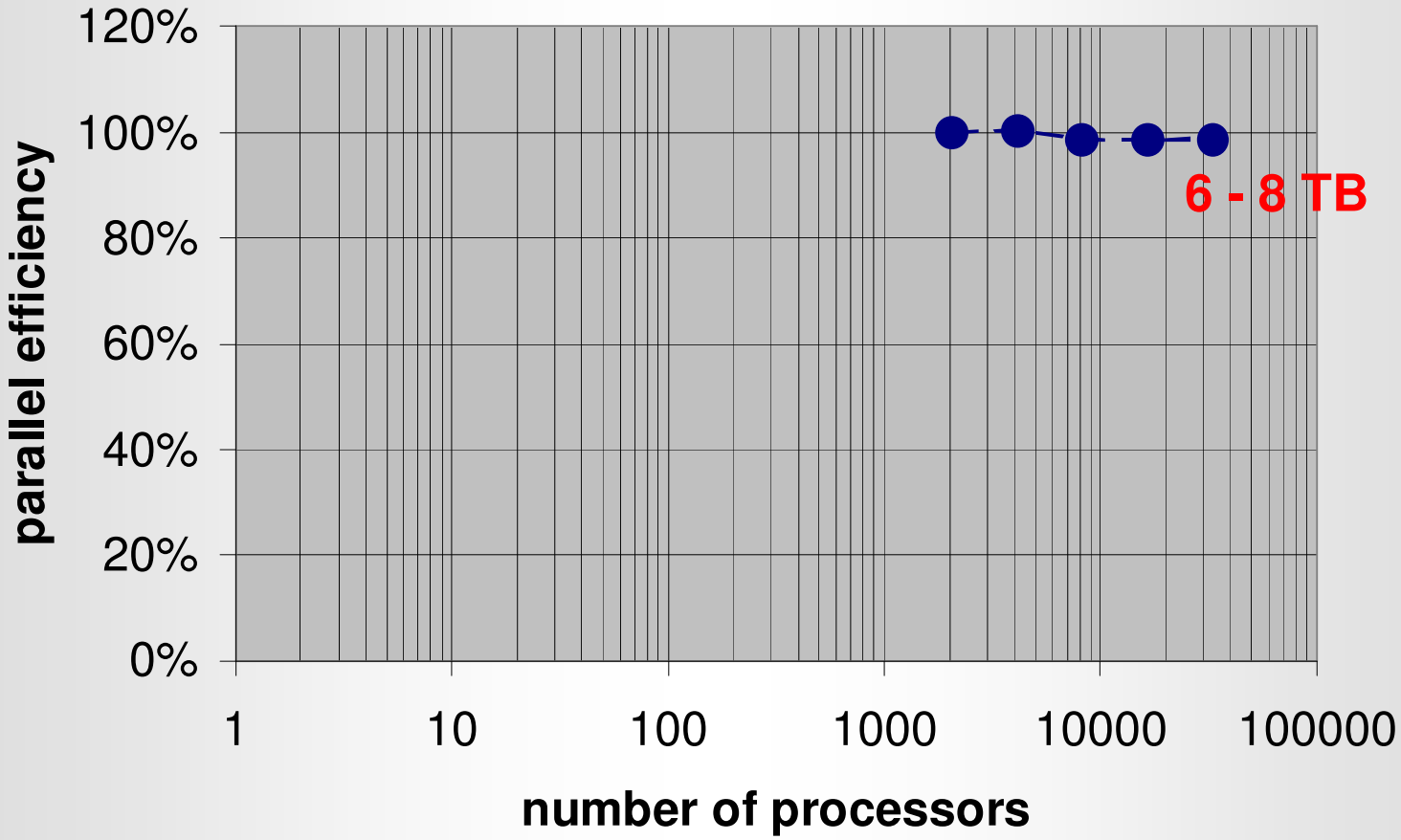
GENE v11+ on IBM BlueGene/L



Weak scaling of GENEv11+ normalized to 2k processors

(problem ~200 MB/proc; measurements in virtual node mode at IBM Watson Research Center)

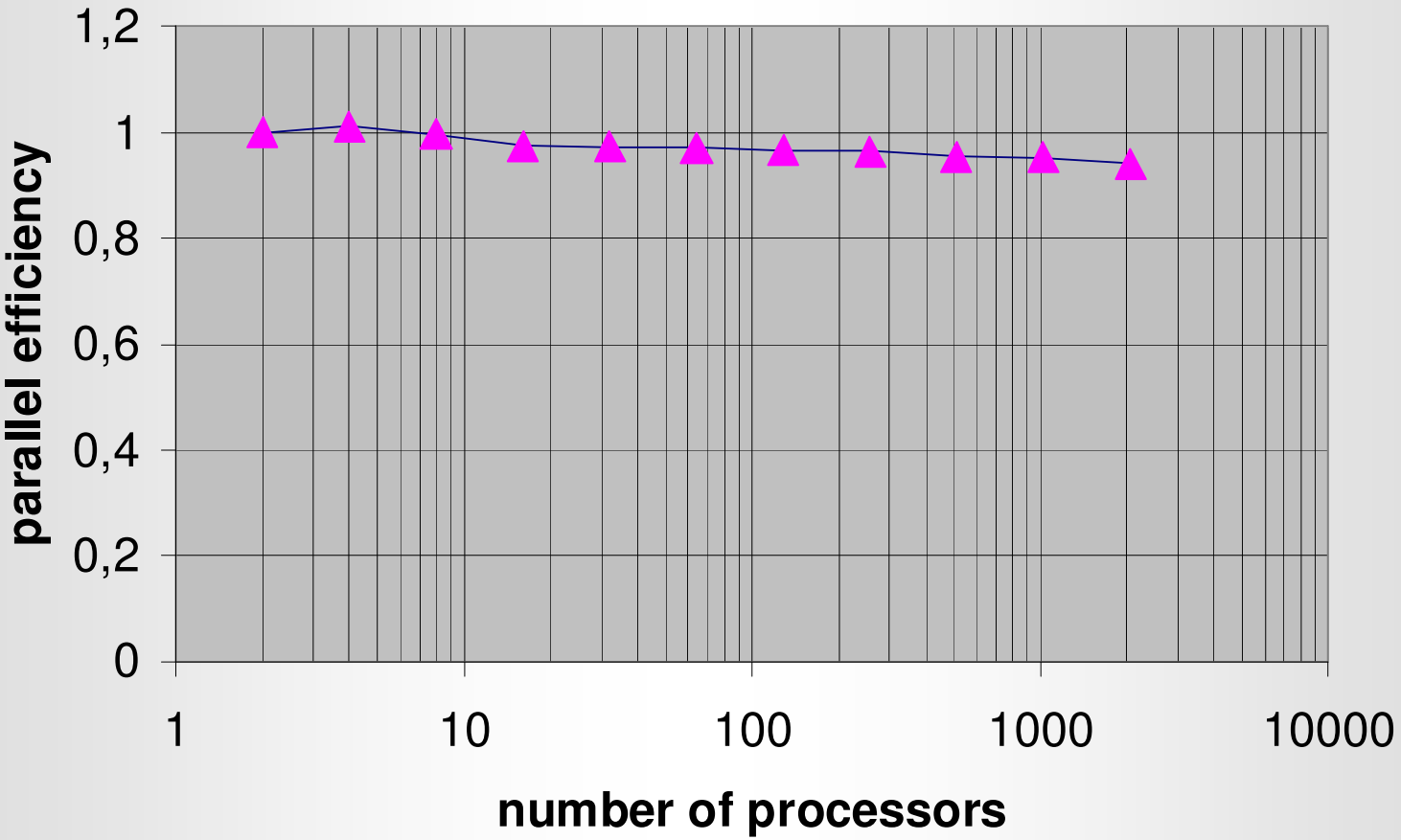
GENE v11+ on IBM BlueGene/L



Weak scaling of GENEv11+ normalized to 2k processors

(problem ~200 MB/proc; measurements in virtual node mode at IBM Watson Research Center)

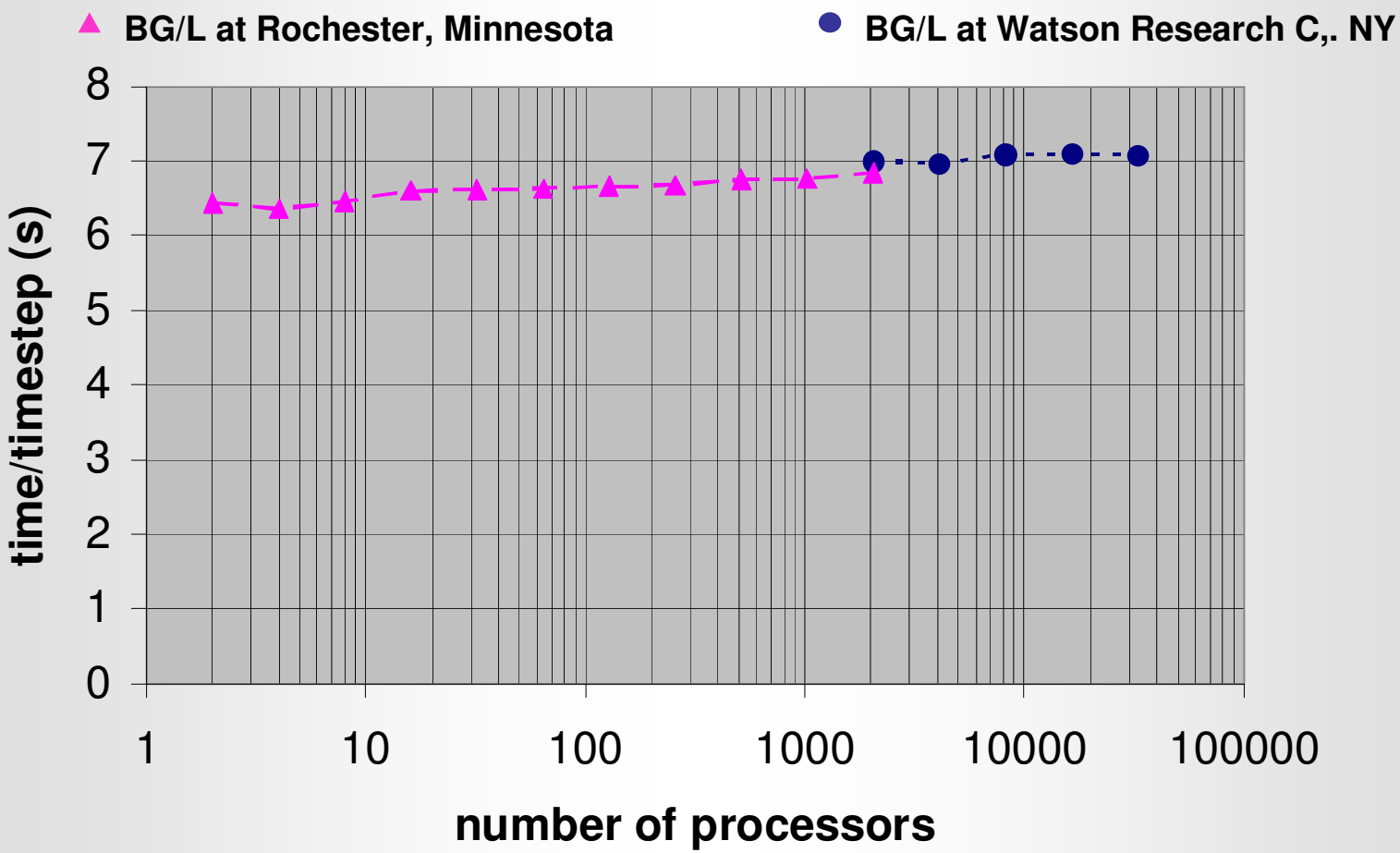
GENE v11+ on IBM BlueGene/L



Weak scaling of GENEv11+ normalized to 2k processors

(problem ~200 MB/proc; measurements in virtual node mode at IBM Rochester, MN)

GENE v11+ on IBM BlueGene/L

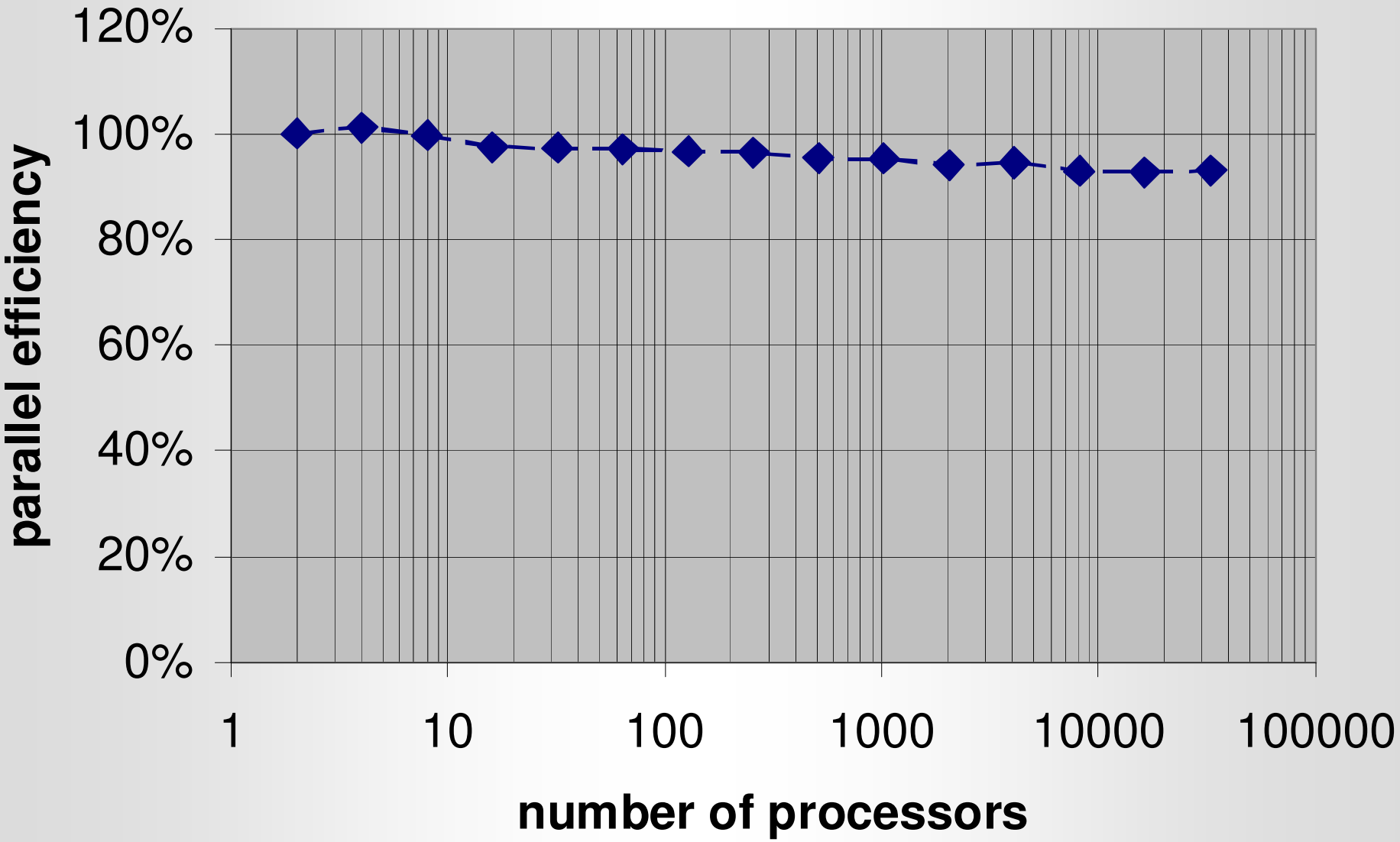


Weak scaling of GENEv11

(problem ~200 MB/proc; measurements in virtual node mode)

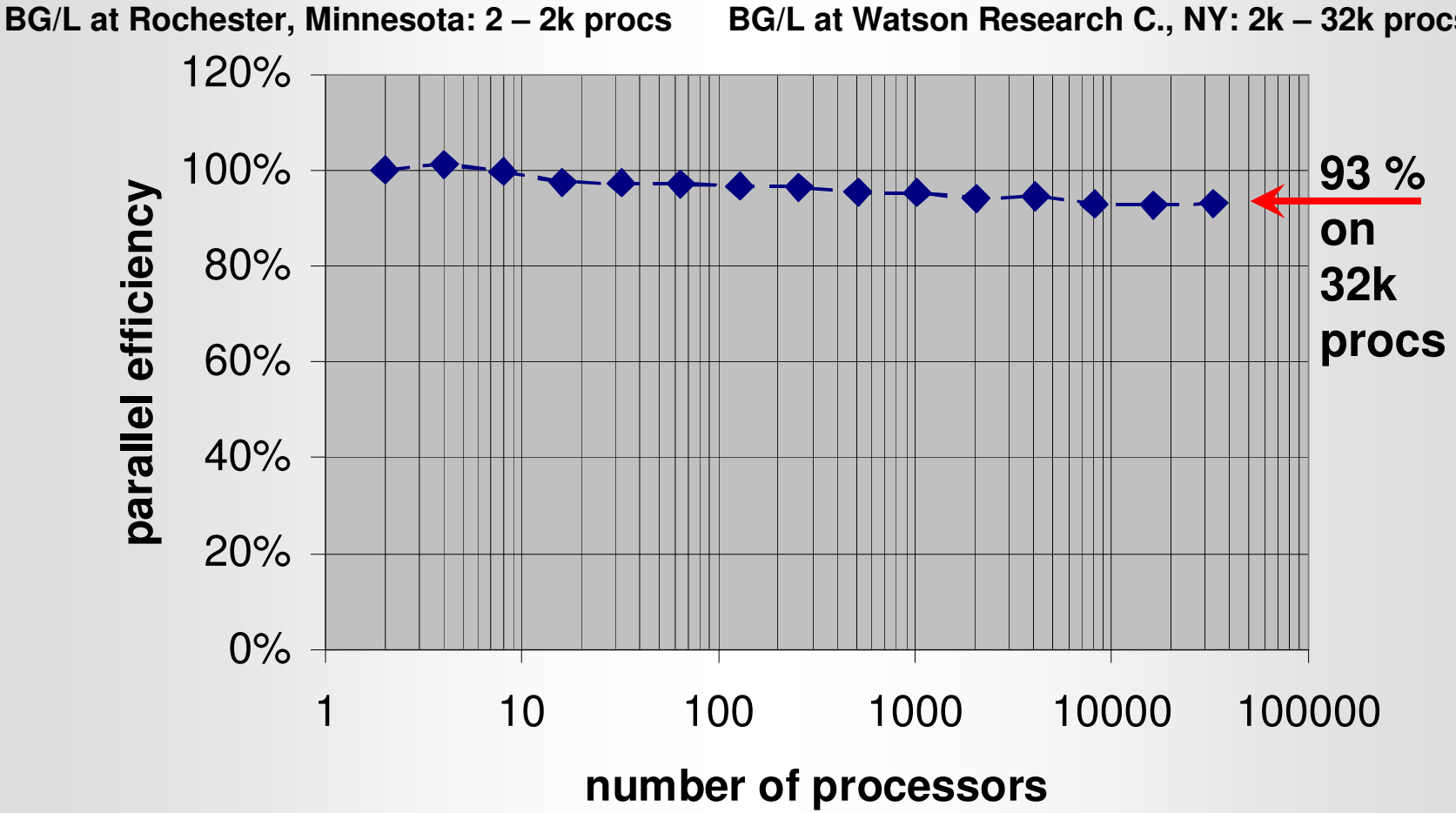
Weak scaling of GENEv11+

15 measuring points / 4 orders of magnitude



RZG IPP

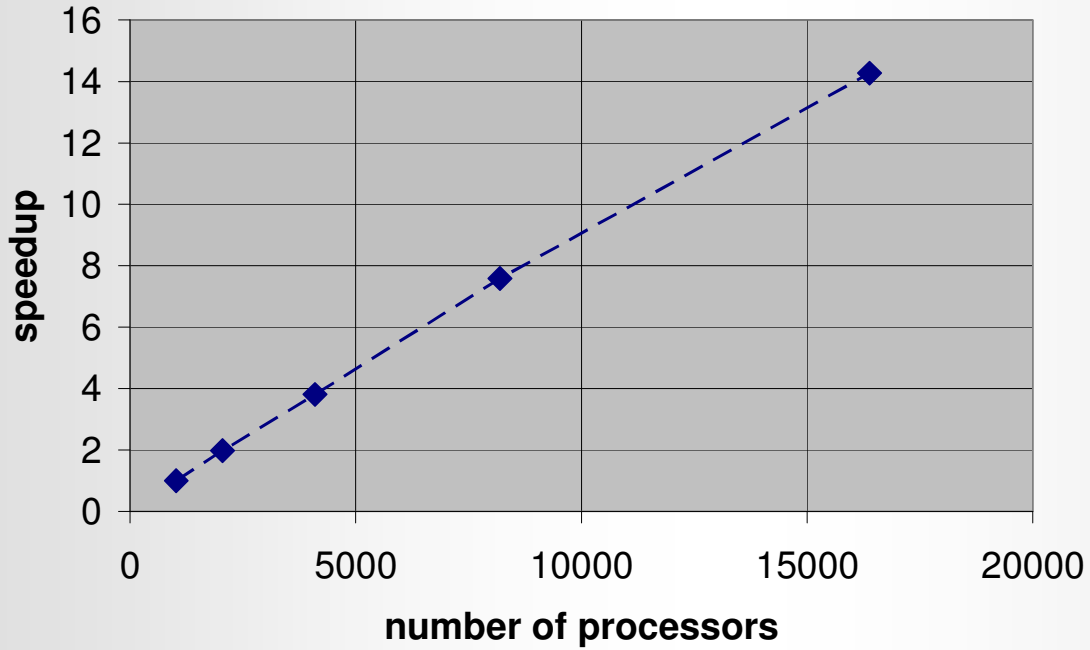
GENE v11+ on IBM BlueGene/L



Weak scaling of GENEv11+ (15 points covering 4 orders of magnitude)
(problem ~200 MB/proc; measurements in virtual node mode, **normalized to 2 processors**)

GENE v11+ on IBM BlueGene/L

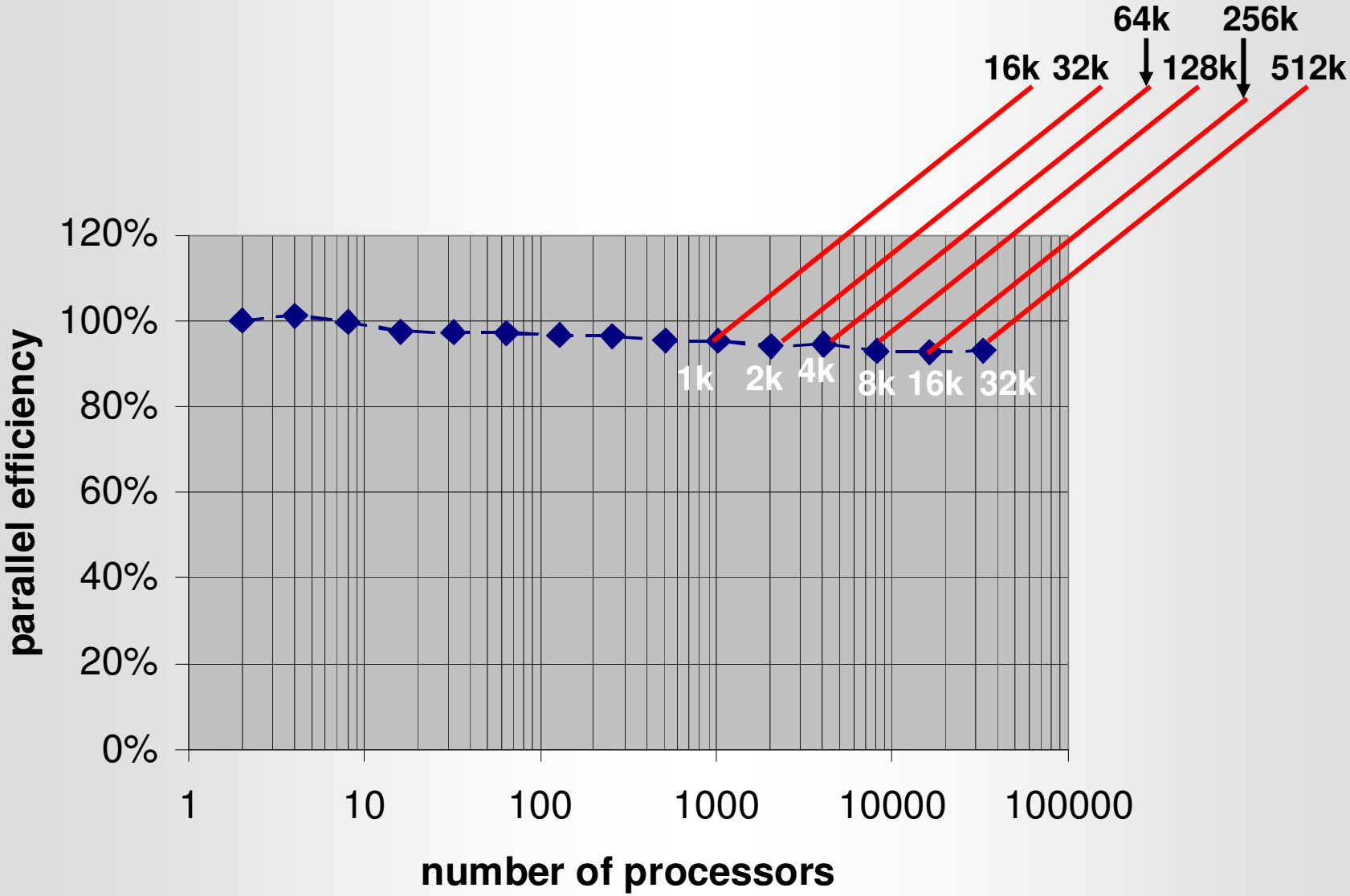
Excellent scalability proven up to 16 times the number of processors used for the base run



Strong scaling of GENEv11+ normalized to 1k processors

(problem ~300-500 GB; measurements in co-processor mode at IBM Watson Research Center)

GENE v11+ on IBM BlueGene/L



Hypothesis

An 8 TB sized GENE problem
can be expected to run efficiently
on 512k processor-cores
of a scalable Petaflops architecture

Question

Are 512k processor-cores the end of the GENE story?

Our assumption is: No!

Why? Parallelization of y dimension not yet exploited!

Why not? Scaling on BG/L not working very well!

Scaling of y dimension on BG/L

BG/L - IBM Rochester, April 2007

GENE11+ 512 GB problem
(nx0=64 nky0=32 ns0=128 nz=64 totalnw=32 nspec=2)

n_pes	npw	npz	nps	npy	time/timestep (s)
2048	1	8	32	4	14.44
2048	16	16	1	4	14.01
2048	32	8	1	4	13.52
2048	1	16	32	2	11.18
2048	16	16	2	2	10.37
2048	32	2	8	2	9.39
2048	32	16	2	1	7.54
2048	32	4	8	1	6.92
2048	32	2	16	1	6.84

Scaling of y dimension on p690+

Power4@1.7 GHz + HPS, JUMP @ FZJ - June 2007

GENE11+ 512 GB problem

n_pes	npw	npz	nps	npy	t/tstep (s)
512	1	8	32	1	9.79
512	1	1	32	8	7.98
512	32	4	2	1	6.93
512	1	1	8	32	6.91
512	16	8	2	1	6.66
512	1	8	1	32	6.60
512	16	2	8	1	6.45
512	1	2	4	32	6.40
512	2	8	8	2	6.40
512	1	4	2	32	6.36
512	32	1	8	1	6.34
512	2	1	4	32	6.28
512	4	8	4	2	6.26
512	1	2	8	16	6.24
512	4	4	8	2	6.20
512	1	8	2	16	6.19
512	2	8	4	4	6.17
512	4	2	1	32	6.08

n_pes	npw	npz	nps	npy	t/tstep (s)
512	4	8	2	4	6.06
512	2	4	1	32	6.05
512	32	2	2	2	6.02
512	8	1	1	32	5.97
512	16	4	2	2	5.94
512	2	4	2	16	5.92
512	16	2	4	2	5.90
512	4	1	4	16	5.82
512	32	1	4	2	5.79
512	4	2	2	16	5.73
512	32	1	2	4	5.68
512	8	1	2	16	5.64
512	8	4	2	4	5.61
512	16	1	1	16	5.58
512	16	2	2	4	5.58
512	8	2	2	8	5.57
512	32	1	1	8	5.47

RZG IPP

GENE11+ scaling of y dimension: high sensitivity for interconnect!

512 GB problem

measured on a further HPC system without network topology optimization

n_pes	npw	npz	nps	npy	t/tstep (s)
512	2	2	2	32	159.82
skip 8 entries					
512	1	8	1	32	153.06
512	4	1	4	16	75.21
skip 13 entries					
512	2	4	2	16	69.27
512	1	1	32	8	42.35
skip 19 entries					
512	32	1	1	8	38.36
512	1	2	32	4	18.47
skip 23 entries					
512	32	1	2	4	15.70
512	4	1	32	2	11.76
skip 26 entries					
512	16	8	1	2	9.46
512	8	1	32	1	9.38
skip 22 entries					
512	32	2	4	1	7.88

RZG IPP

Conclusions

Two important European
plasma turbulence simulation codes,
GENE and ORB5,

could be optimized and adapted to very high scalability

as a major step towards efficient usage
of forth-coming new generations
of petaflops supercomputers

required for realistic simulations of ITER.

Acknowledgements

We thank IBM for access to the BlueGene/L systems
at Watson Research Center,
at Rochester Center (Minnesota),
and J. Pichlmeier for support on using the systems.

We thank the European Commission for support
through contracts FP6-508830 and FP6-031513.